CLAIMS

What is claimed is:

1. A method for processing data representing documents, comprising:

for individual documents of a set of documents, executing a software program to obtain a list of terms found in each document;

comparing the list of terms for a first document to the list of terms for a second document; and

declaring the first document to be substantially identical to, or substantially similar to, the second document if some predetermined number of terms are found in each of the lists of the first document and the second document.

2. A method as in claim 1, wherein if the predetermined number is about 90% of the terms or greater the first document is declared to be substantially identical to the second document.

3. A method as in claim 1, wherein the set of documents is obtained in response to a search query made to a data communications network.

4. A method as in claim 1, and further comprising storing the lists of terms in a database.

5. A method as in claim 1, and further comprising computing a signature for each document, and storing the computed document signature.

6. A method as in claim 1, and further comprising computing a signature for each document, and storing the computed document signature in association with the list of terms for each document.

7. A method as in claim 1, wherein the step of executing a software program assigns to each term a collection-level importance ranking or Information Quotient (IQ), and wherein the IQ is considered during the step of comparing.

8. A method as in claim 1, wherein the step of comparing includes a preliminary step of sorting the documents into a document list in order of increasing size, and where the step of comparing compares a given document with the next larger documents in the document list.

9. A method as in claim 1, wherein the step of comparing includes a preliminary step of sorting the documents into a document list in order of increasing size, and where the step of comparing compares a given document only with another document in the list that is no more than a predetermined amount larger than the given document.

10. A method for processing data representing documents, comprising:

for individual ones of documents, executing a software program to obtain a list of terms found in each document;

computing a document signature for each document from the list of terms obtained for the document;

comparing the document signature for a first document to the document signature for a second document; and

declaring the first document to be substantially identical to the second document if the document signatures are substantially equal.

11. A method as in claim 10, wherein the step of computing a document signature computes a hash code for each term of the list of terms, and then sums all of the hash codes to form the document signature.

12. A method as in claim 10, wherein the documents are obtained in response to a search query made to a data communications network, and where the steps of comparing and declaring are executed in substantially real time as the documents are returned by the query.

13. A method as in claim 10, wherein the documents are obtained in response to a search

query made to a data communications network, where the steps of comparing and declaring are executed in substantially real time as the documents are received from the data communications network, and for a case where a received document is found to be substantially identical to an already received document, returning only one of the documents in response to the search query.

14. A method as in claim 10, and further comprising storing the computed document signatures in a database.

15. A method as in claim 10, and further comprising storing the computed document signature in association with the list of terms for each document..

16. A system for processing data representing documents comprising, for individual documents of a set of documents, a processor for executing a software program to obtain a list of terms found in each document and for comparing the list of terms for a first document to the list of terms for a second document, said processor being operable for declaring the first document to be substantially identical to, or substantially similar to, the second document if some predetermined number of terms are found in each of the lists of the first document and the second document.

17. A system as in claim 16, wherein if the predetermined number is about 90% of the terms or greater the first document is declared to be substantially identical to the second document.

18. A system as in claim 16, wherein the set of documents is obtained in response to a search query made to a data communications network.

19. A system as in claim 16, and further comprising a memory containing a database for storing the lists of terms.

20. A system as in claim 16, said processor being further operable for computing a signature for each document, and further comprising a memory for storing the computed document signature.

21. A system as in claim 16, said processor being further operable for computing a signature for each document, and further comprising a memory for storing the computed document signature in association with the list of terms for each document..

22. A system as in claim 16, wherein said processor, when executing the software program, assigns to each term a collection-level importance ranking or Information Quotient (IQ), and wherein the IQ is considered by said processor when comparing the list of terms for the first document to the list of terms for the second document

23. A system as in claim 16, wherein said processor is further operable, before comparing the lists of terms, to sort the documents into a document list in order of increasing size, and to then compare a given document with the next larger documents in the document list.

24. A system as in claim 16, wherein said processor is further operable, before comparing the lists of terms, to sort the documents into a document list in order of increasing size, and to then compare a given document only with another document in the list that is no more than a predetermined amount larger than the given document.

25. A system for processing data representing documents, comprising, for individual documents of a set of documents, a processor for executing a software program to obtain a list of terms found in each document, for computing a document signature for each document from the list of terms obtained for the document; for comparing the document signature for a first document to the document signature for a second document; and for declaring the first document to be substantially identical to the second document if the document signatures are equal.

26. A system as in claim 25, wherein said processor computes the document signature by computing a hash code for each term of the list of terms, and summing all of the hash codes to form the document signature.

27. A system as in claim 25, wherein the documents are obtained in response to a search query made to a data communications network, and where the processor executes the comparing and declaring functions in substantially real time as the documents are returned

by the query.

28. A system as in claim 25, wherein the documents are obtained in response to a search query made to a data communications network, where the processor executes comparing and declaring functions in substantially real time as the documents are received from the data communications network, and for a case where a received document is found to be substantially identical to an already received document, said processor returns only one of the documents in response to the search query.

29. A system as in claim 25, and further comprising a memory containing a database for storing the computed document signatures.

30. A system as in claim 29, and further comprising storing the computed document signature in association with the list of terms for each document.

31. A computer program recorded on a computer-readable media, said computer program comprising instructions for directing a data processor to process data representing documents by, for individual documents of a set of documents, obtaining a list of terms found in each document; comparing the list of terms for a first document to the list of terms for a second document; and declaring the first document to be substantially identical to, or substantially similar to, the second document if some predetermined number of terms are found in each of the lists of the first document and the second document.

32. A computer program recorded on a computer-readable media, said computer program comprising instructions for directing a data processor to process data representing documents by, for individual ones of documents, obtaining a list of terms found in each document; computing a document signature for each document from the list of terms obtained for the document; comparing the document signature for a first document to the document signature for a second document; and declaring the first document to be substantially identical to the second document if the document signatures are equal.

33. A computer program as in claim 32, wherein the document signature is computed by computing a hash code for each term of the list of terms, and summing together all of the

hash codes to form the document signature.